A Dataset for Action Recognition in the Wild*

Alexander Gabriel[1], Serhan Coşar[1], Nicola Bellotto[1], and Paul Baxter[1]

Lincoln Centre For Autonomous Systems (L-CAS),
University of Lincoln, United Kingdom
{agabriel,scosar,nbellotto,pbaxter}@lincoln.ac.uk
https://lcas.lincoln.ac.uk

Abstract. The development of autonomous robots for agriculture depends on a successful approach to recognize user needs as well as datasets reflecting the characteristics of the domain. Available datasets for 3D Action Recognition generally feature controlled lighting and framing while recording subjects from the front. They mostly reflect good recording conditions and therefore fail to account for the highly variable conditions the robot would have to work with in the field, e.g. when providing in-field logistic support for human fruit pickers as in our scenario. Existing work on Intention Recognition mostly labels plans or actions as intentions, but neither of those fully capture the extend of human intent. In this work, we argue for a holistic view on human Intention Recognition and propose a set of recording conditions, gestures and behaviors that better reflect the environment and conditions an agricultural robot might find itself in. We demonstrate the utility of the dataset by means of evaluating two human detection methods: bounding boxes and skeleton extraction.

Keywords: Agricultural Robotics · Dataset · Human-Robot Interaction · Intention Recognition · Action Recognition

1 Introduction

An agricultural robot has to co-operate with human field workers efficiently, comfortably, and safely. To perform in this setting, the robot needs to understand the intentions behind worker behaviour and basic communication: with a desire to maintain reliability in challenging environments, gestures form an ideal medium.

Developing an Intention Recognition (IR) system for an autonomous, agricultural robot, requires a dataset relevant to task and domain. The agricultural setting differs substantially from the settings present in many existing datasets and the disambiguation challenges inherent to IR suggest a selection of behaviors, which includes actions that cannot be uniquely matched to an intention without taking into account additional information. This takes the task beyond Action Recognition (AR).

* Supported by the RASberry project (https://rasberryproject.com), and the CTP for Fruit Crop Research (www.ctp-fcr.org).

Existing datasets for 3D AR are often recorded under optimized lighting conditions, exhibit few if any artifacts, and show the subject well framed and conveniently oriented. An autonomous robot operating around the year in a field cannot rely on such consistent conditions and so our algorithms must be able to perform well in a less-optimized environment. The dataset we present in this work models such suboptimal conditions. Recordings at multiple distances, a natural background, changing weather, and a variety of clothing styles combine to model a wide range of detection challenges.

The term *intention* is used in various ways throughout the literature, which can be characterized in three main groups. In the first, an intention is synonymous with an action i.e., a series of movements with an atomic purpose like e.g. picking up a glass or turning right [7,27]. In the second, an intention is synonymous with a wish [23] or demand [22], command [13], intended meaning [26,5], goal [12,11,23], plan [16,23], or a goal-plan pair [14], where a plan is a series of actions that changes the state of the environment to a goal state. In the third group, an intention is the meaning [8] of, explanation [28] for, or idea [32] behind an action, plan or utterance.

We see our work as consistent with this third group. Our position is that intentions are not Actions, and that they can neither be observed directly nor unequivocally inferred from movement. The same movement might be performed with different intents, e.g. someone might rub their hands to either warm or clean them or to put on some lotion. The same intent might also lead to different movements being performed, e.g. two people congratulating each other might fist-bump, high-five, or shake hands.

Intentions can also not be directly equated to Plans or Goals, as doing this disregards the possibility that the same Plan could be followed to the same Goal in service of different higher-level intentions. Taking into account contextual information should enable an understanding of other agents that surpasses Plan or Goal Recognition. We call the task of discerning intentions in the face of these ambiguities *Intention Recognition*.

Fig. 1. Left: robot (SAGA Robotics Thorvald II) in front of our poly-tunnels. **Middle:** sensor setup used in the recording. **Right:** robot collecting crates of fruit. **Bottom:** experimental setup: an actor performing actions and behaviors at various distances from the robot.

The dataset we propose contains behaviors that exhibit these ambiguities, e.g. the act of pointing at something is inherently ambiguous in terms of intent. Although it is clear, that the pointing individual wants to draw attention to something in some direction, both the specific target as well as the reason for why they want to draw attention to it might be far from obvious.

The main contributions of this work are twofold. Firstly, we provide a new dataset for outdoor people and action detection[1], which is recorded by a robot in a setting consistent with our agricultural target domain (see Figure 1). Secondly, we propose a methodology for the creation of such datasets, which takes into account the specific features of robot, task, and environment.

In Section 2 we will give an overview of related work before introducing the dataset in Section 3 and demonstrating two applications in Section 4.

2 Related Work

Existing datasets for AR come in basically two varieties: smaller, purposefully recorded datasets featuring good framing, lighting and often multiple sensors on the one side and significantly larger datasets, collected from YouTube and therefore limited to 2D video, but featuring a large variety of recording conditions, subjects, and action classes on the other side. Tables 1 and 2 give an overview of popular datasets. *Classes* gives the number of different action classes a dataset consists of, *Reps* refers to the number of samples collected per class and subject.

Popular datasets for AR from 3D joints or depth video all fall into the first category. Action classes found in this category mostly consist of basic movement, basic interaction with objects [30,31] (picking up, dropping, tossing) and people [33,4] (hugging, kissing, shaking hands, punching, kicking) but also include domain specific classes, e.g. personal hygiene [24], eating/drinking [24], donning/doffing clothes and accessories [29,24], office-style interactions (reading, writing, using laptops, phones) [24], and Wii-like menu navigation and gaming [19,9,3]. The action topic has a significant influence on the subset of actions covered by a dataset. Sensor data provided usually includes 2–3 takes of RGB+D video and 3D-joint positions as produced by the Microsoft Kinect v1 or v2, the NTU RGB+D [24] dataset additionally provides infrared video. A few datasets in this class (e.g. NTU RGB+D [24], Northwestern-UCLA [29]) feature simultaneous recordings from multiple viewpoints.

Although most of the work in this area has focussed on full-body skeletons and various human activities, there are also datasets with a special focus on communicative gestures like we require for our use case. Examples of this are the MSR Gesture 3D [18] dataset, which provides 12 categories of gestures from the American Sign Language (ASL), and the MSRC-12 Kinect [9] dataset, which consists of 12 gestures for interaction with a video game console.

In the pursuit of high-quality data, work in this category generally has tried to optimize recording conditions like the background, illumination, location, and

[1] The dataset is available upon request at `https://lcas.lincoln.ac.uk/wp/research/data-sets-software/outdoor-action-intention-recognition-dataset-rasberry/`

Dataset	Classes	Subjects	Reps	Topics
MSR Action 3D[19]	20	10	2–3	gaming, general movement
MSR DailyActivity[30]	16	16	2	general movement, object interaction
MSR Gesture 3D[18]	12	10	2–3	sign language symbols
MSRC-12 Kinect[9]	12	30	4–5	gaming
SBU dataset[33]	8	21 pairs	1–2	human and item interaction
UTKinect-Action[31]	10	10	2	general movement, object interaction
Northwestern-UCLA[29]	10	10	4–5	clothes, general movement, objects
UTD-MHAD[3]	27	8	4	Wii-like menu navigation, gaming
L-CAS 3D Social[4]	8	10 pairs	1	social interaction
NTU RGB+D[24]	60	40	2	clothes, food, general movement, gestures, human interaction, hygiene

Table 1. Overview of 3D AR Datasets

Dataset	Classes	Samples	Topics
HMDB-51[17]	51	6,849 videos	face actions, human interaction, general movement, sports
UCF-101[25]	101	13,320 videos	cooking, hobby, hygiene, housework, sports, musical instruments
MPII Human Pose[1]	410	25,000 images	hobby, household, hygiene, occupations, musical instruments, sports, transportation
Sports-1M[15]	487	1,133,158 videos	sports

Table 2. Overview of 2D AR Datasets

the distance to the subject. This leads to 3D-joint trajectories close to the actual movement but limits the transferability of algorithmic results to settings like ours, which differ significantly from these optimized conditions.

The second group of datasets was created for AR from 2D videos, generally using Deep Neural Networks. As they don't require multimodal data, researchers can make use of publicly available video sources like YouTube and therefore achieve a much wider variety of subjects and recording conditions, as well as larger numbers of samples. Subjects in these datasets are recorded at different distances and angles, as well as under a great variety of lighting conditions. The datasets additionally feature a generally larger number of action classes (51–487).

Samples of general movement and human interaction can be found in the HMDB-51[17] dataset. More specialized action classes like occupations, hobbies, personal hygiene, playing a wide variety of musical instruments, or using various kinds of transport—all can be found in the MPII Human Pose[1] and UCF-101[25] datasets where the former focusses on good framing in a mix of outdoor and indoor environments and features special face-related actions (e.g. smiling, smoking, talking, chewing), while the latter sports big variation in recording conditions. Even more specialized is the Sports-1M[15] dataset, featuring 487 different classes of sportive activities.

Despite the wide variety of action classes in 2D AR datasets, there are still no action classes fitting the agricultural domain and the limitation to 2D data is a definite downside in a setting where changing illumination, weather and background make the combination of multiple different sensors highly beneficial.

3 The Dataset

The construction of the dataset was guided by two principles. First, a set of discrete actions was chosen to be consistent with the purpose and application domain of the dataset. This incorporated both actions directly related to the activities undertaken by human fruit-pickers and actions with communicative and interactive intent relevant to the domain. The robot needs to interpret both types correctly: activities (such as walking and crate manipulation) are important clues as to the state of human co-workers, gestures (to ensure robot approach and retreat for example) may be interpreted as commands. Consistent with the definition of intentions we have committed to above, these gestures (in particular pointing) do not necessarily correspond to a unique underlying intention.

Second, a structure for the recording process was established, such that a range of aspects could be systematically characterized. This included the use of multiple subjects performing the same set of actions over the same defined set of distances from the recording robot.

The application of these two principles provides a dataset creation methodology that produced an annotated set of ground-truthed, discrete actions, relevant to the agricultural application domain. The dataset can form the basis for evaluation and testing of human and action/intention recognition algorithms, as we demonstrate below.

3.1 Features

The dataset was recorded on a piece of grassland, under varying lighting conditions (sunny, cloudy, morning to afternoon) and at distances ranging from 5m to 50m. The robot used for recording was equipped with a range of sensors that produced data for the dataset, including RGB+D and thermal images, and 3D LIDAR. We recorded 10 actors, performing every activity once at each distance. All participants provided written informed consent, with followed ethical approval from the University of Lincoln College of Science Research Ethics Committee (approval ID: CoSREC459). Behaviors were performed facing away

Fig. 2. Examples of varying light and weather conditions during the recording of the dataset (top). And an indication of resolution of a human at increasing distance from the camera, highlighting the human detection problem over longer distances (bottom).

from the robot, facing to the side and towards the robot for a basic coverage of different directions. A list of activities can be found in Table 3. An overview of distances is shown in Figure 2. After recording, each frame up to 25m distance was labeled with distance, actor ID, action and the direction the actor was facing. Labeling at further distances was hampered due to the actor being too small in the frame (see Figure 2).

Gestures and Activities Being able to detect different behaviors allows the robot to learn a model of the activities of each individual worker which allows it to predict the timing of future support requests. We chose a range of behaviors observable from human fruit-pickers at work, and a set of gestures we deem helpful for basic communication over distances between 10 and 50 meters. To be able to direct the robot's attention to the worker in need of support, we selected a waving and a pointing gesture. For comfortable and efficient loading of the robot, we want to direct it to a preferred stopping distance. To this end, we selected the beckoning, stop and shoo gestures. For basic feedback purposes, we further included a thumbs-up/down gesture and a variant using the lower arm instead of the thumb, which should be easier to detect at greater distances.

Activity	Duration [s]	Description
wave	3.73	With the upper arm
come	2.20	stretched out to the
stop	2.25	side or front,
shoo	2.22	the subject performs
thumb up	1.71	the respective motion
thumb down	1.90	with their hand.
arm up	1.92	
arm down	2.09	
point 0°	1.92	The subject stretches
point 45°	1.91	their upper arm
point 90°	2.00	out to the front
point 135°	1.82	with fist clenched
point 180°	1.81	except for the
point 225°	1.88	index finger which
point 270°	1.99	is also outstretched.
point 315°	1.63	
crate up side	1.30	The subject is standing up
crate down side	1.21	or crouching down while
crate up toward	1.11	holding a crate. They are
crate down toward	1.34	facing either away or
crate up away	1.29	towards the camera,
crate down away	1.83	or to the side.
walk away	3.20	
walk away (crate)	2.20	

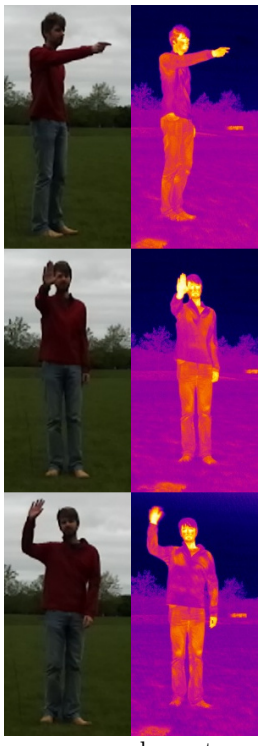

Table 3. Table: characteristics of the set of actions. **Images:** example gestures recorded with RGB (left) and thermal (right) cameras.

The choice of activities is inspired by our application, the collection of fruit crates from human field workers and transportation of said crates to a cooling facility outside the field. The most common activities in this domain are (besides the picking of berries) walking and turning around, crouching down, and standing up—each of these activities occurs with free hands and while carrying a crate.

Table 3 shows the average duration for each action and behavior, as well as their descriptions. The individual actions have relatively short (<4s) durations and many of them (such as waving, shooing or the 'come here' gesture) consist of many, much shorter, movements. A system running motion-based AR on this dataset will have to perform at a challengingly high framerate in order to capture these movements.

3.2 Characteristics

Recording the dataset outside, resulted in a number of characteristics differentiating it from indoor datasets. Outside, the robot and its sensors are subjected to influences from the environment in a variety of ways. In sunny weather, the robot accumulates heat due to the sun's radiation. Insects are then attracted to this new heat source and while buzzing around the robot, fly across the field of view of its sensors (see Figure 3). This results in artifacts that might influence detection algorithms.

But there is not only sunny weather, our recording encompasses stable sunny and cloudy conditions as well as rapidly changing cloud cover (see Figure 2). These influence brightness, hue, saturation as well as the harshness of shadows present in RGB video recordings. Another weather factor is wind, which will distort the recorded body shape either due to its effect on loose clothing or hair. Higher wind speeds than those present during our recordings can force humans to adapt their movement to the horizontal force being applied to them and thus significantly changes movement patterns.

Fig. 3. Occlusions due to environmental conditions; in this case flies close to the robot.

Range further has a significant influence on the performance of detection algorithms, since subjects further away are captured at a lower resolution (see Figure 2). To capture this effect we recorded data from 5m to 50m, which reduces the number of pixels per subject by the square of the magnification factor.

For the characterization we combined the hand-gesture classes (wave, come, stop, shoo, thumb-up, thumb-down) into a single class (hand_gesture), as the skeleton models we use do not support explicit hand detection. Detection of individual fingers at longer distances is ultimately complicated by limited sensor resolution.

While the dataset does not contain occlusions from objects in the environment, it does contain self-occlusions and those due to the presence of the crate. As this does not entirely reflect the nature of agricultural environments (where plants and other agricultural equipment may reduce visibility), this is an area for addressing in further data collection efforts.

4 Applications

In order to demonstrate the utility of our dataset, we use it to evaluate two methods for human detection: bounding boxes, and skeleton extraction. In both cases, established algorithms are applied to RGB images from the dataset, with performance evaluated. The systematic recording methodology used facilitates a rigorous characterization of performance in both cases.

4.1 Skeleton Extraction

We tested extracting 2D skeletons from RGB images[2], up to 25m. The skeletons were extracted using a deep-learning based multi-person skeleton extractor called OpenPose [2].

The average confidence score for skeleton extraction shown in Figure 4 are averages over the confidence scores produced by OpenPose for each skeleton over the duration of an action at various distances. OpenPose returns confidence scores between 0 and 1.

The data shows significantly better skeleton extraction for action classes where the actor is facing the camera (wave, hand and arm gestures) compared to classes where the actor is facing to the side or away (crate actions, pointing) for a part of the sample set. This results from self-occlusion of the further body side occurring in the side views and self-occlusion of the arms when the actor is performing some action while facing away from the camera. As expected, the extraction confidence progressively diminishes with distance, as the number of pixels covering the subject grows smaller (see Figure 2).

[2] As well as false-color thermal images. For results on these, and a more detailed comparison with the RGB results, see [10]

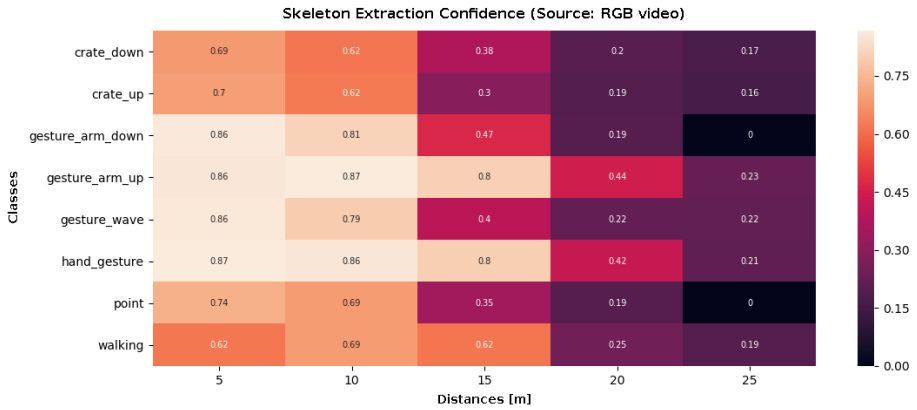

Fig. 4. Average Skeleton detection confidence from ZED RGB+D camera sensor (single RGB video) source on the right. Distances on the X-axis from 5m to 25m, confidence values ranging from 0 to 1. Notable here is the expected loss of extraction confidence with increasing distance. For a more detailed analysis and comparison to skeletons extracted from thermal false-color video, refer to [10].

4.2 Bounding Box Fitting

We have also tested extracting 2D bounding boxes from RGB images to identify humans. We used a deep-learning-based single-shot object detector called YOLOv3 [21]. The detector is run using the pre-trained model for the COCO dataset [20]. Figure 5 shows some examples of person detection using YOLOv3. We evaluated the performance of the person detector by running on 800 annotated images from the dataset (see Section 3). Following the PASCAL Visual Object Classes Challenge [6], the precision and recall rates are calculated by assuming a correct detection, if the area of overlap a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} exceeds 50%.

Figure 6 presents the Precision-Recall curves for various poses and distances, respectively. We can see that person detection works best when people walk

Fig. 5. Examples of person detection whilst performing actions with crates, using the YOLOv3 algorithm.

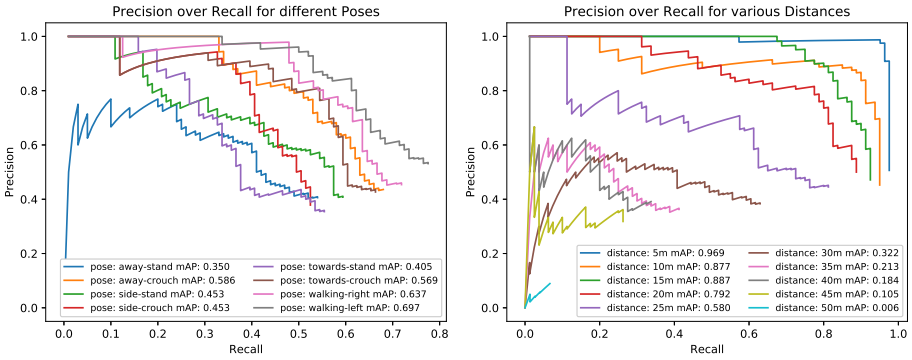

Fig. 6. Precision-Recall curves of person detector for various poses and distances.

laterally. The worst performance is obtained when people face away from the robot. We can also see that after 25m, the algorithm fails to detect people, highlighting an area for prospective improvement in outdoor environments.

5 Conclusion

Our case studies (see Section 4) show that bounding box extraction is less susceptible to large distances compared to skeleton extraction but ultimately fails as well. It performs best when the subject is walking laterally and worst when the subject is standing. Subject orientation does not have an influence on the performance of bounding box extraction, it does however affect skeleton extraction. This is to be expected as facing away from the camera occludes individual joints for the gestures and behaviors we have chosen. For a more detailed analysis of the skeleton extraction and a comparison with skeletons extracted from false-color images recorded by a thermal camera, see [10]. The results of our experiments together suggest that the use of multiple different sensors has the potential to achieve more robust detection performance over a greater variety of conditions. This demonstrates the utility of the systematic approach to the dataset creation that we have employed here.

The choice of actions included in the dataset is motivated by a dual emphasis on AR and IR (hence the hand-gestures, pointing, etc.) and the agricultural context (hence the crate actions). While in this paper we have focused on the human detection aspects in outdoor environments, that this dataset lends itself to, for our wider efforts towards IR (in the context of safe and effective agricultural Human-Robot Interaction) this dataset provides a first step to the ability of characterizing human behavior estimation algorithms, and brings us closer to our goal of appropriately shaping the robot's behavior in response.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition **2017-Januar**, 1302–1310 (2017)
3. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proc. of the IEEE International Conference on Image Processing. pp. 168–172 (Sep 2015)
4. Coppola, C., Faria, D., Nunes, U., Bellotto, N.: Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5055–5061 (2016)
5. Elzer, S., Carberry, S., Zukerman, I., Chester, D., Green, N., Demir, S.: A probabilistic framework for recognizing intention in information graphics. In: Proc. of the International Joint Conference on Artificial Intelligence. pp. 1042–1047 (2005)
6. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (Jan 2015)
7. Fernandez, V., Balaguer, C., Blanco, D., Salichs, M.A.: Active human-mobile manipulator cooperation through intention recognition. In: Proceedings of the 2001 IEEE International Conference on Robotics and Automation. pp. 2668–2673 (2001)
8. Fleischman, M., Roy, D.: Why Verbs are Harder to Learn than Nouns: Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning. In: Proc. of the Annual Meeting of the Cognitive Science Society (2005)
9. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1737–1746. CHI '12, ACM, New York, NY, USA (2012)
10. Gabriel, A., Bellotto, N., Baxter, P.: Towards a Dataset of Activities for Action Recognition on Open Fields. In: to be published in the Proceedings of the UKRAS 2019 Conference on Embedded Intelligence (2019)
11. Giersich, M., Kirste, T.: Effects of Agendas on Model-based Intention Inference of Cooperative Teams. In: Proc. of the International Conference on Collaborative Computing: Networking, Applications and Worksharing. pp. 456–463 (2007)
12. Haigh, K.Z., Geib, C.W., Miller, C.A., Phelps, J., Wagner, T.: Agents for recognizing and responding to the behaviour of an elder. In: Proceedings of the AAAI-02 Workshop Automation as Caregiver. pp. 31–38 (2002)
13. Iba, S., Paredis, C.J.J., Khosla, P.K.: Intention Aware Interactive Multi-Modal Robot Programming. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (2003)
14. Kanno, T., Nakata, K., Furuta, K.: A method for team intention inference. International Journal of Human-Computer Studies **58**(4), 393–413 (apr 2003)
15. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
16. Kiefer, P., Stein, K.: A framework for mobile intention recognition in spatially structured environments. In: Proceedings of the 2008 Workshop on Behaviour Monitoring and Interpretation. pp. 28–41 (2008)

17. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
18. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: 2012 Proceedings of the 20th European signal processing conference (EUSIPCO). pp. 1975–1979. IEEE (2012)
19. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. pp. 9–14 (June 2010)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv (2018)
22. Schrempf, O.C., Hanebeck, U.D.: A Generic Model for Estimating User Intentions in Human-Robot Cooperation. In: Proceedings of the 2nd International Conference on Informatics in Control, Automation and Robotics. vol. 3, pp. 251–256 (2005)
23. Schrempf, O.C., Schmid, A.J., Hanebeck, U.D., Wörn, H.: A Novel Approach To Proactive Human-Robot Cooperation. In: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication. pp. 555–560 (2005)
24. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
25. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild (November) (2012)
26. Sykes, E.R., Franek, F.: Inside the Java Intelligent Tutoring System Prototype: Parsing Student Code Submissions with Intent Recognition Edward. In: Proceedings of the 2004 International Conference on Computers and Advanced Technology in Education. pp. 613–618 (2004)
27. Tahboub, K.: Intention recognition of a human commanding a mobile robot. In: Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems. vol. 2, pp. 896–901 (2004)
28. Tahboub, K.A.: Intelligent human-machine interaction based on Dynamic Bayesian Networks probabilistic intention recognition. Journal of Intelligent and Robotic Systems: Theory and Applications **45**(1), 31–52 (2006)
29. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.: Cross-view action modeling, learning, and recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2649–2656 (June 2014)
30. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1290–1297. IEEE (2012)
31. Xia, L., Chen, C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–27 (June 2012)
32. Youn, S., Oh, K.: Intention recognition using a graph representation. International Journal of Applied Science, Engineering and Technology **4**(1), 13–18 (2007)
33. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE (2012)